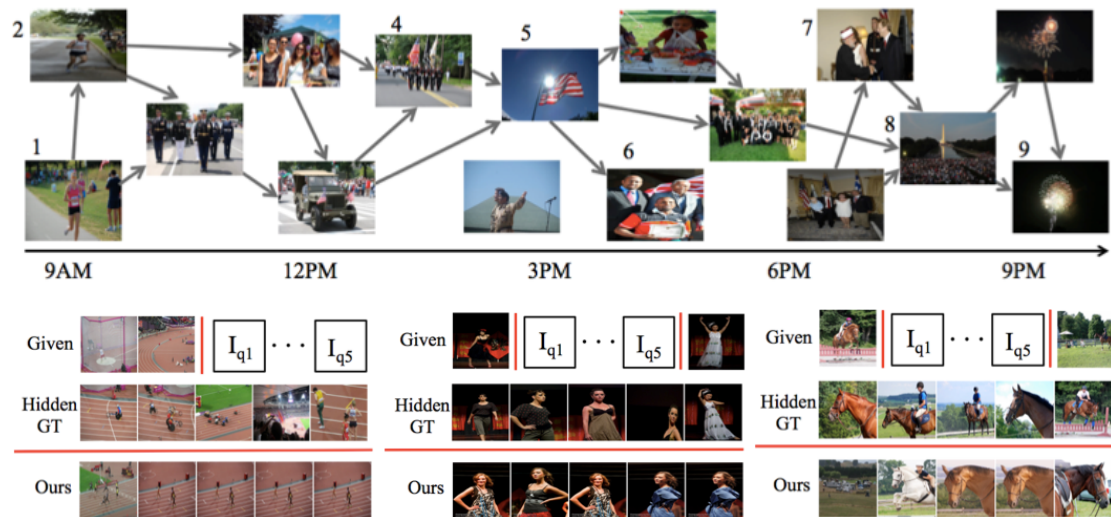


## Weekly Report (2014.09.9~09.14)

### Done

1) 先把手头上正在进行的一些任务简单总结下:

a) 贝叶斯网络: 暂时没有什么新的进展, 文章 “Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos” 对 flick 的图片用动态贝叶斯网络进行训练, 生成 stroyline graph, 可以根据已有的图片对缺失的图片进行预测。可以参考, 对出警数据是否也可以进行类似的预测。



b) 七巧板针对个人的多源异构数据的分析

数据维度有: 居民的电话、住址、工作单位、民航、旅馆、加油、卡口、上网、停车、信用卡等。

初步设想, 仅围绕车, 可以做以下方向:

- 有车一族出行规律的分析: 这个数据比基站的数据更为精确, 且有家庭、工作单位住址加以辅助, 更适合做这个事情。
- 交通拥堵原因(红绿灯、测速、事故、道路瓶颈等)分析:  
今年 vis 北大中的文章 “Visual Exploration of Sparse Traffic Trajectory Data” 比较侧重可视设计, 可以分析道路拥堵的规律(比如周末和工作日拥堵时间分布的差异), 但并未对拥堵原因作进一步研究, 只是一笔带过。《预知社会》那本书里介绍了一些拥堵的模型可以借鉴, 或许可以通过可视化的手段(对交通密度、交通流率等进行可是分析), 帮助用户发现拥堵的原因。

c) 多源异构时空数据的检索

这个问题目前还没找到很好的点, 唯一确定的是 time-space cube 可以尝试用 psh 的方法进行压缩, 快速索引。但是如果牵涉到原始数据, 后来想想主要瓶颈还是在磁盘读取的速度上, 怎么合理存储磁盘上得原始数据, 使得每次 query 的平均磁盘访问次数减少比较重要, 而这个应该是和具体的查询任务相关。所以这块还没找到好的突破点。

2) 关于 Data cube 的压缩, 也看了一些相关工作, 但都是针对数值型的数据。

最早的有用小波的方法进行压缩(具体没看)。还有核密度估计的方法, 直接对 cube 进行密度估计, 查询的时候可以快速地返回一个近似的结果。之后的 condensed cube, 基于 base single tuple 进行压缩, 基于 data cube 的稀疏性, 可以无损地进行压缩。再之后的 Dwarf, 提出了 Prefix redundancy 和 Suffix redundancy, 通过对两者进行去冗余达到压缩的目的, 方法和 nanocube 的基本一样, 可以看做是 nanocube 的前身。04 年还有一篇针对高维(100

维级别)的数据,用倒排索引的方式进行压缩。

- 3) 气象方面,上周完成了数据的下载、拼合。这周加上的传输函数、数据参数的处理,可以在前端选择数据类型及传输函数。

## **To Do**

- 1) 我准备看 PSH 和 datacube 的代码,一方面之后可能要用,另一方面也有助于更好理解方法的适用场合。
- 2) 并行方面将数据的下载添加到并行后端,可以达到如果数据缺失自动下载的效果。